Kapitel 2 Grundlagen diagnostischer Verfahren

Abschnitt 2.1 Allgemeines zu psychologischen Tests

Was sind zentrale Elemente der Definition von Psychologischen Tests?

Für einen psychologischen Test gilt: a) Es handelt sich um eine Messmethode, bei der Personen auf standardisierte Reizvorlagen (Aufgaben, Fragen etc.) reagieren. b) Reaktionen werden durch die spezifischen, im Test realisierten Bedingungen hervorgerufen. c) Die Reaktionen erlauben einen wissenschaftlich begründbaren Rückschluss auf die individuelle Ausprägung eines psychologischen Merkmals (oder auch mehrere Merkmale). d) Das Vorgehen ist standardisiert. e) Ziel ist eine quantitative (Ausprägung des Merkmals) und/oder eine qualitative Aussage (Vorhandensein oder Art des Merkmals) über das psychologische Merkmal. (a bis c zitiert nach Heidenreich 1993, S. 389)

Was versteht man unter einer Testtheorie?

Testtheorien machen grundlegende Annahmen darüber, wie aufgrund von beobachtetem Verhalten in einem Test auf ein latentes Merkmal geschlossen werden kann. Sie dienen als Grundlage für die Konstruktion und Evaluation von psychologischen Tests.

Was versteht man unter reflexiven Messungen?

Bei reflexiven Messungen geht man davon aus, dass Antworten auf Testitems die Ausprägung der Testpersonen in dem zu messenden latenten Merkmal reflektieren. Im Gegensatz dazu gehen formative Messungen davon aus, dass nicht die zugrunde liegende latente Variable auf die Itemantworten „einwirkt“, sondern die Itemantworten bilden die latente Variable.

Abschnitt 2.2 Die Klassische Testtheorie

Was sind zentrale Annahmen der Klassischen Testtheorie?

Eine zentrale Annahme der Klassischen Testtheorie ist, dass Messungen fehlerbehaftet sind. Sie nimmt an, dass eine einzelne Messung aufgrund von unsystematischen Einflussfaktoren ein höheres oder niedrigeres Ergebnis liefert als aufgrund der tatsächlichen Merkmalsausprägung zu erwarten wäre. Basierend darauf nimmt die Klassische Testtheorie an: Es gibt einen wahren Wert der Merkmalsausprägung einer Person v, definiert als der Erwartungswert unendlich häufiger Messungen unter identischen Bedingungen (Existenzaxiom). Sowie: Der beobachtete Wert einer Person v in einem Testitem i setzt sich zusammen aus dem wahren Wert der Person v und einem zufälligen Messfehler (Verknüpfungsaxiom). Es folgt weiterhin aus der Annahme zufälliger Messfehler, dass diese mit wahren Werten, beobachteten Werten oder Messfehlern aus anderen Messungen unkorreliert sind – sonst wären sie nicht unsystematisch bzw. zufällig.

Was versteht man in der Klassischen Testtheorie unter dem „wahren Wert“?

Die Klassische Testtheorie versteht unter dem „wahren Wert“ den Erwartungswert aller Messergebnisse (in einem Test oder von Testwiederholungen).

Wie lässt sich (zur Reliabilitätsschätzung) bemessen, wie hoch der Anteil der Varianz der wahren Werte und wie hoch der Anteil der messfehlerbedingten Varianz ist?

Im Kern nutzt man dazu die Korrelation des Tests mit einer (essenziell) parallelen Version des Tests oder die Korrelation zwischen (essenziell) parallelen Testteilen bzw. Testwiederholungen.

Was wird an der Klassischen Testtheorie kritisiert?

Es wird kritisiert, dass sich Messfehler nicht, wie in der Klassischen Testtheorie postuliert, immer zufällig um den wahren Wert verteilen. Zudem wird kritisiert, dass die Parameter der Klassischen Testtheorie populations- und stichprobenabhängig sind. Zudem geht man in der Klassischen Testtheorie davon aus, dass sich bei häufiger Wiederholung von Messungen, etwa durch mehrere parallele Items, Messfehler „herausmitteln“. Damit geht man auch davon aus, dass Ergebnisse von einzelnen Messungen aggregiert werden können – etwa zu einem Summen- oder Mittelwert. Dies setzt jedoch mindestens ein Intervallskalenniveau und die Prüfung, ob Summen- und Mittelwerte angemessene Berechnungen des Testwertes sind, voraus. Eine explizite Prüfung dessen bleibt jedoch in der Praxis häufig aus.

Abschnitt 2.3 Item-Response-Theorien

Nennen Sie verschiedene Item-Response-Modelle und die Antwortformate, für die diese verwendet werden können!

Für dichotome Daten und das Ziel, Merkmale zu quantifizieren, bieten sich dichotome Raschmodelle an. Für dichotome Daten und das Ziel, eine Klassifikation von Personen vorzunehmen, bieten sich latente Klassenanalysen an. Mixed-Rasch-Modelle erlauben eine Kombination aus Quantifikation und Klassifikation. Für ordinale Daten kann das ordinale Rasch-Modell herangezogen werden, um Merkmale zu quantifizieren. Eine Klassifikation auf Basis ordinaler Daten gelingt mit einer Klassenanalyse für ordinale Daten. Nominale Daten können mit Rasch-Modellen bzw. Klassenanalysen für nominale Daten ausgewertet werden.

Was versteht man unter einer itemcharakteristischen Kurve?

Unter einer itemcharakteristischen Kurve versteht man den Verlauf der Lösungswahrscheinlichkeit eines Items in Abhängigkeit von der Fähigkeit der Testpersonen.

Wie ist die Itemschwierigkeit im einparametrischen dichotomen Rasch-Modell definiert?

Die Itemschwierigkeit ist definiert als der Punkt auf dem Merkmalskontinuum, an dem die Lösungswahrscheinlichkeit $p(X_{vi} = 1) = .50$ beträgt.

Was versteht man unter spezifischer Objektivität?

Spezifische Objektivität der Vergleiche ist eine Eigenschaft des dichotomen einparametrischen Rasch-Modells. Sie meint, dass zur Ermittlung des Fähigkeitsunterschieds zwischen Personen es egal ist, welche Items herangezogen werden. Spezifische Objektivität der Vergleiche meint auch, dass Vergleiche von Items nicht davon abhängen, welche Personen man dazu heranzieht.

Was versteht man unter lokaler stochastischer Unabhängigkeit?

Lokale stochastische Unabhängigkeit beschreibt die Annahme, dass bei konstanter zu messender Eigenschaft bzw. Fähigkeit die Lösungswahrscheinlichkeit eines Items unabhängig von dem Ergebnis bei einem anderen Item sein sollte.

Erläutern Sie die Logik des grafischen Modelltests!

Eine einfache Möglichkeit, zu prüfen, ob für einen vorliegenden Test tatsächlich von spezifischer Objektivität der Vergleiche ausgegangen werden kann, bietet der grafische Modelltest. Bei diesem teilt man die Stichprobe in 2 Substichproben (z. B. ältere und jüngere Personen), berechnet die Itemschwierigkeiten getrennt für beide Stichproben und inspiziert dann anhand eines Streudiagramms, ob die getrennt geschätzten Schwierigkeitsparameter der Items konvergieren. Dies sieht man im Streudiagramm daran, dass sich die Schwierigkeiten größtenteils entlang einer Winkelhalbierenden anordnen.

Welche Erweiterungen des einparametrischen dichotomen Rasch-Modells gibt es und wodurch unterscheiden diese sich von selbigem?

Im 2PL-Modell wird neben einem Personenfähigkeits- und einem Itemschwierigkeitsparameter auch ein Diskriminationsparameter angenommen. Dadurch verlaufen Item-charakteristische Kurven nicht mehr notwendigerweise parallel, sondern können unterschiedlich steil sein. Die vorteilhaften Eigenschaften des 1PL-Modells der spezifischen Objektivität der Vergleiche sowie des Summenwerts als erschöpfende Statistik treffen auf das 2PL-Modell nicht zu. Das 3PL-Modell enthält zusätzlich einen Rateparameter.

Was beschreiben Item- und Testinformationsfunktion?

Die Iteminformationsfunktion beschreibt die Bereiche auf dem Merkmalskontinuum, für die ein bestimmtes Item besonders „informativ“ ist. Aggregiert man die Iteminformationsfunktionen aller Items eines Tests, so erhält man die Testinformationsfunktion. Sie gibt an, für welche Bereich auf dem Merkmalskontinuum ein Test (mit seinen Items) besonders informativ ist.

Welche Schwellenabstände impliziert das Ratingskalenmodell?

Das Ratingskalenmodell geht davon aus, dass die Schwellenabstände innerhalb der Items unterschiedlich, aber über die Items hinweg gleich sind.

Wozu dient eine Latent-Class-Analyse?

In manchen Fällen besteht das Ziel einer Testung nicht darin, die Merkmalsausprägung von Personen zu quantifizieren, sondern darin, Personen verschiedenen Klassen zuzuordnen. Hierzu kann die Latent-Class-Analyse genutzt werden. Damit können Gruppen identifiziert werden, die sich bezüglich der Antworten auf die applizierten Items unterscheiden. Eine zentrale Annahme ist, dass sich zwischen den Gruppen die Lösungswahrscheinlichkeiten der Items unterscheiden. Innerhalb einer Gruppe erhält jedes Item eine Lösungswahrscheinlichkeit, die für alle Personen der Gruppe gleich ist.

Abschnitt 2.4 Konstruktionsprinzipien psychologischer Tests

Nennen Sie wichtige Schritte des Testentwicklungsprozesses!

Wichtige Schritte des Testentwicklungsprozesses sind: Festlegung des Ziels eines Messinstruments; Definition des Messgegenstands; Itemgenerierung; Itemanalyse und Analyse des Testentwurfs; Revision des Itempools; erneute Itemanalyse und Analyse des Testentwurfs; Validierung. Ggf. können weitere Schritte (im Anschluss oder zwischendurch) notwendig sein.

Nennen und beschreiben Sie grundlegende Methoden der Itemgenerierung!

Geht es um die Messung eines Merkmals, stellt die deduktive Methode (häufig auch als rationale Methode bezeichnet) für viele Testentwicklerinnen und -entwickler die ideale Lösung dar. Man beruft sich auf eine Theorie, die eine gute Beschreibung des Merkmals liefert, und formuliert Items entsprechend der theoretischen Vorgaben. Bei der induktiven Methode der Testentwicklung stützen sich die Personen, die einen Test konstruieren, nicht primär auf eine bestimmte Theorie. Die Strategie besteht vielmehr darin, diejenigen Items zu einem Testentwurf zusammenzufassen, die hoch miteinander korrelieren und damit (sehr wahrscheinlich) gemeinsam ein latentes Merkmal abbilden. Die kriteriumsorientierte Methode kommt zum Einsatz, wenn statt der Position einer Person in Relation zu einer Vergleichsnorm das Erreichen oder Verfehlen eines konkreten Kriteriums ermittelt werden soll. Items werden

so formuliert und ausgewählt, dass sie für das Kriterium repräsentativ sind. Ansatzpunkt der externalen Methode der Testentwicklung ist das Vorliegen verschiedener Gegebenheiten in der Realität. Dies können Gruppen von Personen als Teil der sozialen Realität sein. Dazu werden Items ausgewählt, die zwischen Mitgliedern und Nicht-Mitgliedern dieser Gruppen differenzieren oder mit relevanten externen Gegebenheiten korrelieren.

Welche Randbedingungen sind bei der Itemformulierung zu beachten?

Als Randbedingungen bei der Itemformulierung sind zu beachten: Zielgruppe (wer soll den Test später bearbeiten?); Anwendungsbereich (welche Verwendungszwecke sind angedacht?); Einsatzbedingungen (unter welchen Randbedingungen wird ein Test eingesetzt? Und von wem?).

Was sind die Vor- und Nachteile gängiger Antwortformate?

Völlig freie Antworten sind geeignet, wenn komplexes Denken, originelle Lösungen oder Praxistransfer erfasst werden sollen, ohne dass die Lösung durch vorgegebene Antwortalternativen eingegrenzt oder vorgebahnt wird. Nachteile bestehen in der aufwendigen Auswertung, der meist eingeschränkten Auswertungsobjektivität sowie dem Umstand, dass Antworten ggf. abhängig von mündlicher bzw. schriftlicher Ausdrucksfähigkeit sind. Eingeschränkt freie Antworten sind geeignet, wenn verfügbares Wissen und nicht bloßes Wiedererkennen erfasst werden soll, ebenso für originelle Lösungen. Nachteilig sind die eher aufwendige Auswertung und eine eventuell eingeschränkte Auswertungsobjektivität. Zuordnungsaufgaben (und Sortieraufgaben) sind zur Erfassung von Wissen und Kenntnissen geeignet und dabei objektiv und ökonomisch (mit Schablone, Auswertungsprogramm) auszuwerten. Es wird jedoch nur das Wiedererkennen und nicht der freie Abruf von Gedächtnisinhalten erfasst. Multiple-Choice-Aufgaben (und Forced-Choice-Aufgaben) sind ebenfalls objektiv und ökonomisch auszuwerten. Dabei sind die Nachteile, dass gute Distraktoren oft schwer zu finden sind, dass bei Leistungstests nur das Wiedererkennen und nicht der freie Abruf von Gedächtnisinhalten erfasst wird sowie dass die richtige Lösung in Leistungstests auch durch Raten erreicht werden kann. Beurteilungsaufgaben sind ebenfalls objektiv und ökonomisch auszuwerten und liefern differenziere Informationen als eine dichotome Antwortskala. Dichotome Antwortformate haben den weiteren Nachteil, dass eine Entscheidung (zwischen zwei Antwortalternativen) erzwungen wird.

Was versteht man unter einer negativen Itempolung?

Von einer negativen Itempolung spricht man, wenn der negative Pol einer Antwortskala für eine hohe Merkmalsausprägung spricht und umgekehrt. Bspw. würde eine hohe Zustimmung (positiver Pol der Antwortskala) bei Items zur Depressivität wie „Ich bin meist fröhlich“ oder „Ich bin selten gedrückter Stimmung“ für eine geringe Depressivität sprechen.

Abschnitt 2.5 Grundzüge von Itemanalysen

Wie ist Itemschwierigkeit in der Klassischen Testtheorie definiert?

Die Itemschwierigkeit nach der Klassischen Testtheorie gibt an, wie groß der Anteil der Personen ist, die das Item im Sinne des Merkmals beantwortet haben. Je höher der Anteil der Personen ist, die ein Item im Sinne des Merkmals beantworten, desto leichter ist das Item.

Was versteht man unter einer part-whole-korrigierten Trennschärfe?

Die Trennschärfe eines Items ist definiert als die Korrelation des Items mit dem Test oder Testteil, zu dem dieses Item gehört. Von einer part-whole-Korrektur spricht man, wenn bei der Berechnung des Testwertes als Summe aller Antworten das jeweilige Item, für das die Trennschärfe bestimmt werden soll, unberücksichtigt bleibt.

Von welchen Faktoren hängt die Trennschärfe ab?

Die Höhe der Trennschärfe hängt von der inhaltlichen Passung des Items, der Verteilungsform von Itemantworten und Testwerten sowie von der Streuung des Items und der Testwerte ab.

Was sind Eigenwerte im Rahmen einer Faktorenanalyse und wie kann man an deren Verlauf die Zahl der zu extrahierenden Faktoren ermitteln?

Der Eigenwert eines Faktors beschreibt den Anteil der Varianz aller in die Analyse einbezogener Items, der durch diesen Faktor erklärt wird. Eine Möglichkeit, die Zahl der zu extrahierenden Faktoren zu bestimmen, besteht in der Inspektion des Verlaufs der Eigenwerte über alle denkbaren Faktoren (von nur einem Faktor bis zu so vielen Faktoren, wie Items vorhanden sind). Weist der Verlauf der Eigenwertlinie einen Abfall von einem zum nächsten Faktor auf, bleibt man besser bei der Zahl der Faktoren, die vor dem Abfall der Eigenwerte gegeben war. Optisch stellt sich ein solcher Abfall als Knick im Eigenwerteverlauf dar.

Was beschreibt der Q-Index?

Für Itemanalysen nach Probabilistischen Testtheorien stehen verschiedene Itemfitmaße zur Verfügung. Eines dieser Fitmaße, der Q-Index, folgt der Logik, dass Items dann gut zum angenommenen Modell passen, wenn sie von den „richtigen“ Personen gelöst wurden (Rost 2004). Das heißt, schwierige Items sollten vornehmlich von Personen mit hoher Fähigkeitsausprägung gelöst worden sein. Items mittlerer Schwierigkeit sollten vornehmlich von Personen mit mittlerer und hoher Fähigkeitsausprägung gelöst worden sein, seltener von Personen mit niedriger Fähigkeitsausprägung. Der Q-Index prüft daher, wie wahrscheinlich

das vorliegende Muster der Antworten (aller Personen bei diesem Item) ist, gegeben die Randsumme des Items.

Abschnitt 2.6 Testgütekriterien

Nennen Sie die 3 Hauptgütekriterien und wesentliche Nebengütekriterien!

Unter den 3 Hauptgütekriterien versteht man Objektivität, Reliabilität und Validität. Wesentliche Nebengütekriterien sind: Normierung, Skalierung, Ökonomie, Nützlichkeit, Unverfälschbarkeit, Zumutbarkeit/Akzeptanz, Fairness.

Welche Formen der Objektivität unterscheidet man?

Man unterscheidet Durchführungs-, Auswertungs- und Interpretationsobjektivität. Durchführungsobjektivität ist dann gegeben, wenn ein Verfahren immer auf die gleiche Weise durchgeführt wird. Auswertungsobjektivität gibt das Ausmaß an, in dem Antworten der Testperson unabhängig von der Person, die den Test auswertet, zu den gleichen Ergebnissen führen. Interpretationsobjektivität ist dann gegeben, wenn klare Aussagen über zulässige Interpretationen sowie Hilfsmittel für die Einordnung von Ergebnissen (z. B. in Form von Normtabellen) vorliegen.

Was sind Methoden der Reliabilitätsschätzung und was ist bei deren Anwendung jeweils zu beachten?

Bei der Schätzung über die Retest-Methode wird der gleiche Test derselben Stichprobe 2× dargeboten. Die Korrelation zwischen den Ergebnissen der beiden Messungen wird als Reliabilitätsschätzung interpretiert. Die Retest-Methode setzt voraus, dass man die gleichen Personen zu einem späteren Zeitpunkt erneut untersuchen kann und die dann durchgeführte Messung als mindestens essenziell parallel zur 1. Messung gelten kann. Eine damit verbundene Schwierigkeit der Retest-Methode besteht darin, das Intervall zwischen den beiden Testungen sinnvoll zu wählen. Bei der Schätzung über die Paralleltest-Methode werden (mindestens essenziell) parallele Versionen eines Tests von ein und derselben Gruppe von Personen bearbeitet. Die Reliabilitätsschätzung ergibt sich aus der Korrelation der beiden parallelen Tests. Der wesentliche Nachteil der Paralleltest-Methode besteht darin, dass für die allermeisten Tests keine parallele Version vorliegt. Bei der Schätzung über die Split-Half- bzw. Testhalbierungsmethode wird das Testergebnis auf Basis von 2 äquivalenten (d. h. mindestens essenziell parallelen) Testhälften berechnet. So erhält man für jede Probandin und jeden Probanden 2 Testwerte. Der ganze Test wird von den Testpersonen zunächst normal bearbeitet; die Aufteilung in Hälften erfolgt erst nach Vorliegen der Ergebnisse. Die Korrelation der Werte aus beiden Testhälften wird – nach einer Korrektur – als Schätzung der Reliabilität verwendet. Mithilfe der Spearman-Brown-Formel kann man für diese künstliche Testhalbierung korrigieren und schätzen, wie hoch die Reliabilität des Tests mit der gesamten Itemzahl wäre.

Voraussetzung für die Anwendung der Spearman-Brown-Formel ist jedoch, dass es sich bei den Testhälften um mindestens essenziell parallele Versionen handelt. Bei der Schätzung über die interne Konsistenz wird der Anteil der gemeinsamen Varianz an der Gesamtvarianz der jeweils relevanten Items berechnet. Es stehen hierzu verschiedene Formeln zur Verfügung. Häufig wird das sog. Cronbachs Alpha berechnet. Es muss jedoch beachtet werden, dass Alpha nur dann eine angemessene Schätzung der Reliabilität gewährleistet, wenn die Testitems mindestens tau-äquivalente Messungen darstellen (Eid und Schmidt 2014).

Welche Formen der Äquivalenz von Messungen gibt es und wodurch unterscheiden sich diese?

Es werden 5 Formen der Äquivalenz von Messungen unterschieden, die unterschiedlich „strenge“ Annahmen machen: Parallele Messungen, essenziell parallele Messungen, tau-äquivalente Messungen, essenziell tau-äquivalente Messungen und tau-kongenerische Messungen. Die fünf unterschiedlich strengen Annahmen zur Äquivalenz von Messungen setzen allesamt voraus: Die Messungen müssen eindimensional sein, also nur ein Merkmal messen – die Messfehler der beider Messungen sind unkorreliert. Sofern von parallelen Messungen im strengen Sinne ausgegangen werden soll, müssen die wahren Werte identisch und die Messungen gleich reliabel sein (gleicher Messfehlereinfluss; in nachfolgender Tabelle ist zusammengefasst, welche Annahmen von den 5 Formen der Äquivalenz von Messungen gemacht werden). Sind beide Messungen gleich reliabel, aber die wahren Werte von Personen in einer der beiden Messungen lediglich um eine Konstante verschoben, so spricht man von essenziell parallelen Messungen. Die verbleibenden 3 Formen der Äquivalenz verlangen nicht, dass Messungen gleich reliabel sein sollen. Werden dennoch identische wahre Werte gemessen, spricht man von tau-äquivalenten Messungen. Sind die Messungen nicht gleich reliabel, aber die wahren Werte bis auf eine Verschiebung um eine Konstante identisch, spricht man von essenziell tau-äquivalenten Messungen. Sind die wahren Werte nicht identisch, aber durch eine lineare Transformation ineinander überführbar, spricht man von tau-kongenerischen Messungen (vgl. Bühner 2021; Eid und Schmidt 2014).

Inwiefern ist McDonalds Omega dem Koeffizient Alpha vorzuziehen?

Es muss beachtet werden, dass Cronbachs Alpha nur dann eine angemessene Schätzung der Reliabilität gewährleistet, wenn die Testitems mindestens tau-äquivalente Messungen darstellen (Eid und Schmidt 2014). Sind Messungen durch Testitems nur tau-kongenerisch, ermöglicht Cronbachs Alpha keine angemessene Schätzung der Reliabilität. Da tau-Äquivalenz der Items selten angenommen werden kann, raten mittlerweile viele Forscherinnen und Forscher von der Nutzung von Cronbachs Alpha ab. In diesen Fällen empfiehlt sich eine Schätzung der Reliabilität anhand von McDonalds Omega (McDonald 1999).

Wie kann die Reliabilität einer Messung in der Einzelfalldiagnostik genutzt werden?

Ein wesentlicher Nutzen der Reliabilität einer Messung für die Einzelfalldiagnostik besteht in der Feststellung der Genauigkeit, mit der ein Individualergebnis ermittelt wurde. Man berichtet dazu Konfidenzintervall.

Wodurch wird die Breite eines Konfidenzintervalls, das man um einen Testwert legt, beeinflusst?

Die Breite eines Konfidenzintervalls, das man um einen Testwert legt, wird beeinflusst durch: die Reliabilität der Messung, die angenommene Sicherheitswahrscheinlichkeit, die Standardabweichung der Messwerte, die Wahl eines ein- vs. zweiseitigen Konfidenzintervalls sowie die Methode der Berechnung (Äquivalenz vs. Regressionsmethode).

Welche Rolle spielt die Reliabilität der Messung beim Vergleich von 2 Testwerten einer Person (etwa vor und nach einer Behandlung)?

Aufgrund des Messfehlers könnten sich die zwei gemessene Werte einer Person zufällig voneinander unterscheiden auch wenn keine tatsächliche Veränderung des Merkmals eingetreten ist. Daher muss man ermitteln, wie groß eine Differenz zwischen 2 Werten (des gleichen Tests, gemessen bei der gleichen Person) sein muss, sodass sie wahrscheinlich nicht alleine durch den Messfehler erklärt werden kann. Dies bezeichnet man als kritische Differenz.

Wofür korrigiert eine doppelte Minderungskorrektur?

Die doppelte Minderungskorrektur korrigiert die Korrelation zwischen zwei Messungen um den Messfehleranteil der beiden Messungen. Eine einfache Minderungskorrektur korrigiert diese Korrelation nur um den Messfehleranteil in einer der beiden Messungen.

Wie ist Validität definiert?

Validität bezeichnet das Ausmaß, in dem Evidenz und Theorie die Interpretation von Testwerten rechtfertigen (AERA et al. 2014).

Anhand welcher Testeigenschaften lassen sich Belege für die Validität von Testwertinterpretationen generieren?

Belege für die Validität von Testwertinterpretationen können generiert werden anhand des Testinhalts, von Antwortprozessen, der Struktur des Tests sowie des Zusammenhangs zu anderen Variablen.

Was versteht man unter einem nomologischen Netz?

Unter einem nomologischen Netz versteht man Annahmen über die Beziehung des Zielkonstrukts zu anderen Konstrukten. Sofern die entsprechenden Messungen die angenommenen Konstrukte adäquat abbilden, sollten die beobachteten Testwerte in einem ähnlichen Zusammenhang stehen, wie er für die Konstrukte angenommen wurde. Lassen sich für einen neu entwickelten Test die so abgeleiteten Zusammenhänge nicht zeigen, misst der Test entweder das Konstrukt nicht (oder nur teilweise) oder die angenommenen Zusammenhänge waren theoretisch nicht gut abgeleitet.

Wie kann eine Multitrait-Multimethod-Analyse im Rahmen der Testvalidierung genutzt werden?

Eine einfache Möglichkeit, Methodeneffekte im Rahmen der Testvalidierung zu berücksichtigen, bietet der sog. „Multitrait-Multimethod-Ansatz“ (MTMM-Ansatz; Campbell und Fiske 1959). Er sieht vor, dass alle in einer Validierungsstudie berücksichtigten Konstrukte möglichst mit mehreren Methoden erfasst werden. Die daraus resultierenden Korrelationen werden dann nach einer einfachen Systematik sortiert und bewertet: Statt nur nach „konvergent“ und „diskriminant“ zu unterscheiden, werden Korrelationen zusätzlich nach „gleiche Methoden“ und „ungleiche Methoden“ sortiert. Im Rahmen einer Korrelationsmatrix können dann bspw. die entsprechenden Korrelationen systematisch verglichen werden.

Was versteht man unter retrograden, konkurrenten und prädiktiven Validitätsbelegen?

Damit wird der Zeitpunkt, zu dem die Validitätsbelege erhoben wurden, bezeichnet. Retrograd bezeichnet den Umstand, dass Validitätsbelege bereits vor der Erhebung des Zielkonstrukts erfasst wurden (bspw. Schulnoten, die bereits lange vor der eigentlichen Testung vorlagen). Konkurrent meint, dass Validitätsbelege parallel zur Messung des Zielkonstrukts erfasst wurden. Prädiktiv meint, dass solche Validitätsbelege zeitlich nach der Messung des Zielkonstrukts erfasst wurden.

Wie beeinflusst die Symmetrie/Asymmetrie von 2 Messungen deren Korrelation?

Symmetrie bzw. Asymmetrie beschreibt, dass Prädiktor (z. B. ein neu entwickelter und zu validierender Test) und Kriterium in Bezug auf den Inhalt und das Generalitätsniveau merh oder weniger gut korrespondieren können. In Fällen deutlicher Asymmetrie ist die zu erwartende Korrelation gemindert.

Was versteht man unter Variabilitäts- bzw. Abweichungsnormen?

Variabilitäts- oder Abweichungsnormen geben an, wie weit eine Person mit ihrer Testleistung unter oder über dem Mittelwert einer Vergleichsgruppe liegt. Die Abweichung jedes einzelnen Messwertes vom Mittelwert der Normgruppe wird dabei in Einheiten der Streuung der Normgruppe ausgedrückt. Als Vergleichsgruppe kann bspw. eine bevölkerungsrepräsentative Stichprobe dienen.

Nennen Sie gängige Normwertskalen sowie deren Mittelwert und Standardabweichung!

Gängige Normwertskalen sind die z-Skala (MW = 0, SD = 1), die IQ-Skala (MW = 100, SD = 15), die Standardwertskala (MW = 100, SD = 10), die T-Wertskala (MW = 50, SD = 10) sowie die Stanine-Skala (MW = 5, SD ≈ 2).

Wann kann ein Test als unfair gegenüber einer oder mehreren Personengruppen bewertet werden?

Es existieren verschiedene Auffassungen darüber, wann ein Test als fair oder unfair angesehen werden sollte. Grundsätzlich gilt, dass Personengruppen aufgrund ihrer ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppenzugehörigkeit nicht systematisch benachteiligt werden dürfen. Die amerikanischen Uniform Guidelines for Employee Selection Procedures sprechen von diskriminierenden bzw. unfairen Auswahlverfahren, wenn die Selektionsrate (also der Anteil aller Ausgewählten an allen Bewerberinnen und Bewerbern) in einer Subgruppe unterhalb von 80 % der Gruppe mit der höchsten Selektionsrate liegt.